RESEARCH STATEMENT

Varun Gupta

Booth School of Business guptav@uchicago.edu
University of Chicago. http://www.varungupta.info

Methodology: My core research agenda is to develop theoretical foundations for design and analysis of algorithms for resource allocation in service and inventory systems, viz.: managing servers in a data center, trailer inventory in freight logistics platforms, inventory in a network of warehouses, staffing of call centers, drivers in a ridesharing system, to name a few. The questions I like to explore are *whether there are nice structual characterizations of the optimal policies*, *how sensitive are the performance metrics to changes in input parameters*, and *how can these policies can be made robust or self-adaptive to changes in the input parameters*?

Most recently, my research has focused on studying resource allocation in service systems through one, or a combination, of the following three lenses:

I. *Control policies for non-stationary systems:* Given the knowledge of the probability distributions generating the input to the system, one can formulate the problem of finding the optimal policy as a Markov Decision Process (MDP). However, the statistical properties of the input processes may change unpredictably over time (e.g., traffic to a call center exhibits a strong time-of-day effect). Further, the curse of dimensionality can render exact solution of MDPs computationally prohibitive. One of my research emphasis is to discover: *can simple distribution-oblivious policies match the performance of optimal policies*, *to what extent can they self-adapt to time-varying demand pattens*, and *how can such simple policies be blended with online learning approaches to improve performance and augmented with safety constraints*? Light-weight policies are desirable because "fewer moving parts" make them intuitive and interpretable to human operators, and hence more likely to be adopted in practice – a phenomenon I experienced first-hand during my stint at Google, and some work done as part of a team making logistics recommendations to India's Chief Economic Advisor for managing the Covid Liquid Medical Oxygen supply chain.

II. *Design in the presence of strategic agents:* Any system once designed does not operate in vacuum, but is used by agents who are interested in maximizing their utilities. Not modeling the incentives of these agents at the design stage can lead to unintended consequences on system performance. Many of my research papers revisit service systems where by enlarging the action space of the users to make decisions such as joining a queue, leaving a queue, or swapping positions in a queue leads to more realistic models and interesting insights. More recently, I have been exploring design dynamic auctions, motivated by truckload marketplaces.

III. *Asymptotic scalings for insights into design and analysis of policies:* Since computing an optimal policy often faces the obstacles of intractability and lack of insights, one often studies an asymptotic regime where the systems become increasingly "larger," yet preserve the essential features of the

original system, much like the Central Limit Theorem. If careful attention is not paid to the scaling process then one may end up drawing incorrect conclusions about the original system. The third lens that I employ in my research is that of engineering *scalings with a purpose* – novel asymptotic scalings tailored with an eye towards the design question we aim to answer.

In subsequent sections I briefly describe a few representative projects that highlight the above lenses. A full list of publications is available at the end of the document.

External Engagement: Having spent time in industry (internships at Microsoft Research, Bell Labs, postdoc at Google), and with 10 years of teaching MBA students the utility of tools like optimization modeling and simulation for decision making, I have strived to actively increase engagement with non-academic organizations. In the past I have interacted with the office of India's Chief Economic Advisor on suggestions for improving the logistics of Liquid Medica Oxygen in the midst of the covid pandemic, C.H. Robinson (a freight platform), and Tillable (a platform connecting owners of agricultural land with farmers). I have an ongoing collaboration with Convoy, a third party logistics platform, where I serve in the role of Faculty Scholar. My projects with Convoy involve rethinking their algorithms for trailer rebalancing and the design of their dynamic auction for matching shipments to carriers.

I. Control Policies for Non-Stationary Systems [8, 11, 16, 25, 27]

I was trained as a queueing theorist during my PhD. Queueing theory is a branch of applied probability where one makes statistical regularity assumptions – such as a Poisson arrival prcocess of requests, and independent and identically distributed request sizes – to provide quite sharp performance analysis of policies as well as to design near optimal policies. Being in a Computer Science department, I was also surrounded by traditional online algorithm researchers who studied competitive analysis under adversarial instances. Being unconvinced of the superiority of either approaches, I have been spent much time thinking of semi-stochastic/semi-adversarial models which yield novel algorithmic insights, as well as algorithms which are robust for such online models.

The first two projects described below started with the goal of designing control algorithms which are oblivious to the distribution and potential non-stationarity in the arrivals to the service system, but still nearly match the performance of the distribution-aware optimal policies. The third project is a recent foray into exploring the interconnections between learning and control for non-stationary systems.

Online Packing and Matching [1, 5]: Online Bin Packing is a fundamental algorithmic problem which shows up in diverse application such as packing Virtual Machines in physical servers, shipping logistics, and remnant inventory scheduling.[1].The abstract form of the problem is as follows: a sequence of T integer-sized items, sampled *i.i.d.* from an unknown distribution arrive as a stream. The system operator has at their disposal an infinite collection of bins of some integer size B. The goal of the system operator is to pack items on arrival feasibly so as to minimize the number of bins used at the end of the horizon. In [5] we propose the first distribution-oblivious bin-packing algorithm that places items

[1] Adelman, Daniel, George L. Nemhauser. 1999. Price-directed control of remnant inventory systems. *Operations Research* 47(6) 889-898.

to greedily/myopically minimize a convex function (penalized Lagrangian) of the state. Our algorithm turns out to have a deep connection to the *Online Mirror Descent* algorithm for convex optimization. We prove that the additive suboptimality of our algorithm after packing T items is $\mathcal{O}(\sqrt{T})$ on any input distribution. Furthermore, using the connection with convex optimization, we prove that our algorithm exhibits small suboptimality gap on many non-*i.i.d.* models of input sequence.

In a follow-up work [1] we show that online bin packing and many other online resource allocation problems such as network revenue management and assemble-to-order systems can be studied in a unified manner by viewing them as instances of online multiway matching with three types of resources – offline, online-queueable, and online-nonqueueable. Here we show that under a crude knowledge of the input distribution given as the optimal basis of the static planning problem, as long as a certain *robustness of optimal basis* condition holds, a simple greedy algorithm obtains bounded regret.

Traffic-oblivious dynamic capacity scaling to balance energy-performance trade-offs [13]: Consider again the bin packing problem described above. Suppose now the bins map to physical servers in a data center, and items map to Virtual Machines (VMs). Unused bins correspond to idle servers which consume energy and should ideally be powered down. However demand for computing resources exhibits a strong diurnal pattern as well some unpredictability. Naively powering down all idle servers can lead to a severe performance penalty in terms of delay because they can not be powered up instantaneously. The key to a good policy for data center energy management is to strike the right balance between conserving energy by not keeping idle servers powered up, and to guarantee small delay by keeping enough servers powered up. In [13] we propose a traffic-oblivious scheme that achieves provably near-optimal performance for large data centers. The algorithm, called DelayedOff has two components: *(i)* Once a server idles, it starts a count down timer and powers down if the timer hits zero and no new tasks have been assigned to the server; and *(ii)* a task assignment scheme that routes a new arrival to the server which idled most recently, so that the idle periods of the servers are either very short (and hence the server stays powered up), or very long (so that energy saved while these servers are powered down compensates for the delay penalty when they power up later). This algorithm led to a patent, and follow up work which has been implemented at Facebook.

Dynamic Regret Minimization for Non-Stationary Linear Dynamical Systems [2]: This project started due to a lingering question on the bin packing project mentioned above: While the distribution-agnostic heuristic performs very well while being robust, an optimal use of the history of the system to learn the input distribution can not lead to any worse performance (since such information can always be ignored). *How should the ideal policy blend the goal of robustness to non-stationary input with the desire to improve performance by learning from historical data?* While there is a lot of work on non-stationary online learning, e.g. in bandit optimization, the kind of problems we are interested in are MDPs. As a first step, we study Linear Dynamical Systems (in particular the Linear Quadratic Regulator) which are one of the simplest MDPs. Here the state x_t evolves according to linear dynamics: $x_{t+1} = A_t x_t + B_t u_t + w_t$ where u_t is the control, and w_t is stochastic noise. When the dynamics matrices A_t, B_t are unknown and non-stationary, but with "small total variation" over the problem horizon (variation $V_T = o(T)$ measured as the sum of Frobenius norm of changes in A_t, B_t), we propose

a control algorithm which achieves the optimal min-max regret rate of $\widetilde{\mathcal{O}}\left(V_T^{2/5}T^{3/5}\right)$ compared to the optimal dynamic policy. A crucial technical ingredient needed to prove this result was a novel bound on estimation error of the Ordinary Least Squares estimator when the parameter to be estimated is non-stationary.

Next steps: The first two examples were success stories of distribution oblivious algorithms, while the last example was an initial step towards using learning for non-stationary control. We are still not close to answering the questions of *What is the optimal way to combine learning and noisy forecasts with robust control? What should be learned – an end-to-end policy, or a forecasting model? How should safety constraints and historical data about the dynamics be incorporated into online learning for control so that costly exploration is minimized?* These questions form the immediate next items on my research agenda.

II. Designing Service Systems in the Presence of Strategic Agents [9, 10, 31, 33, 34]

I describe the most recent of these below.

Menu Design for Bipartite Matching systems [4, 31]: In problems involving allocation of public goods, such as public housing, where the supply of the public good is heterogenous and limited (e.g., housing units in different neighborhoods, or of different sizes), and at the same time the users have different desirability for the goods, a central planner usually wants to find an allocation mechanism that optimizes the two usually competing objectives of efficiency (the goods should be allocated to those who value them the most) and delays (the users should not wait for a long time to be matched). Motivated by this question we study a bipartite matching queueing model where the system designer allows the users to sign up for one of a pre-designed set of bundles of goods (a.k.a., service classes). As the goods become available, they are allocated First-Come-First-Served to the users. The resulting queueing models, known as multi-class multi-server systems, have a rich literature and, owing to the difficulty in their analysis, have been studied in "conventional heavy-traffic" regime. However, this literature usually makes a rather strong assumption called complete resource pooling, under which the delays of all the service classes equalize. In [4] we took the first steps towards analysis of such queueing models in a conventional heavy-traffic scaling where the limiting matching system breaks into multiple complete resource pooling systems, and proposed a computational approach to designing bundles which optimize the trade-off of efficiency and delays when the system designer can dictate which service classes users join. In [31] we look at the same problem but when users decide which service class to join under anticipated delays and likelihoods of getting matched to a server in the service class of their choice. We find that multiple complete resource pooling components endogeneously emerge in equilibrium when the users are strategic and rational utility maximizers. Finally, we show that simplification that results in heavy-traffic regime allows us to encode the problem of finding service menus that maximize a combination of delay and matching reward under the equilibrium joining behavior via a Mixed Integer Linear Program (MILP) for a fairly rich class of menus.

Next steps: There are numerous unanswered questions as follow up to the work described above, such

as queueing systems in overload (with application to organ transplant queues), other notions of fairness such as min-max, and allowing users to reject an offer. However a direction I am most interested in pursuing is to study how these systems can be designed while being robust to the preferences of the agents since often this information is either not known to the system designer, or known imperfectly. Another research project I am pursuing right now is on designing dynamic auctions for truckload marketplaces. Here the interplay of online arrival of shipments, the online bidding process by carriers, hassle cost on the part of carriers for bidding on too many shipments, and risk aversion towards losing submitted bids creates a rich auction design problem.

III. Asymptotic Scalings for Insights into Design of Service Systems [3, 4, 8, 31, 32]

Asymptotic scaling is the umbrella term for the analysis of sequences of successively "larger" models that preserve essential features of the original system that one wants to understand. Two of the biggest success stories in asymptotic analysis of queueing systems are Kingman's analysis of multiserver systems under "conventional heavy traffic" asymptotics where arrival rate approaches the service capacity, and Halfin and Whitt's "square-root staffing rule" or "many-servers heavy-traffic" asymptotics for multi-server systems where the number of servers as well as the total arrival intensity increase simultaneously while their difference grows as the square root of the number of servers. Whereas Kingman's analysis allows accurate prediction of the impact of variability of arrival and service processes on the queueing delays, the Halfin-Whitt rule turns out to be very accurate for optimal staffing under fairly general staffing cost and delay cost functions. My work in this space is driven by the need to develop and study novel asymptotic scalings, that are tailored to the question the system designer aims to answer. I summarize one of my works in this vein.

How good is greedy load balancing? [8, 17] In a multiserver system with a dedicated queue per server, the load balancing problem is to decide, on the arrival of a job, which server should the job be assigned to. Whereas classical load balancing models either assume that the job size distribution obeys the Exponential law, or that the servers employ FCFS scheduling, my research is motivated by computing applications where servers process all the tasks in their queue in parallel akin to Processor Sharing scheduling discipline. A common policy used in practice, as well as studied theoretically is the greedy Shortest-Queue heuristic – which is known to be the optimal policy when job size distribution follows the Exponential law. Two questions arise: *(i)* Is greedy really a good policy when the job-size distribution is not Exponential, or when job sizes are known? *(ii)* Are there simpler heuristics which match the performance of SQ with less communication between the load balancer and the servers?

The SQ load balancing model is notoriously hard to analyze. In [17] we had found experimental evidence towards a "near-insensitivity" phenomenon for SQ load balancing with Processor Sharing servers: The mean completion time of the tasks only seems to depend on the mean of the job size distribution and not the entire distribution. Further, this insensitivity phenomenon disappeared for many other load balancing policies such as routing to server with the least unfinished work, which performed poorly compared to SQ. In fact SQ, which is oblivious to job sizes, performed quite competitively with a naive myopic size-based load balancing rule. *Given that exact analysis of SQ load balancing is intractable,*

what asymptotic regime can faithfully replicate our experimental findings? Neither conventional heavy traffic or Halfin-Whitt regime offered a solution.

In [8] we propose that the Non-Degenerate Slowdown (NDS) regime is a more meaningful regime to study qualitative properties of load balancing policies. The NDS regime is a many-servers heavy-traffic scaling where the number of servers and the total arrival intensity increase simultaneously, but the difference between the two is held constant. Intuitively, under NDS regime, the queueing delays converge to a non-degenerate distribution. Thus, while NDS regime is never a good regime for capacity planning, it is the perfect regime to study the impact of load balancing policies and job size distribution on the queueing delays. We uncover the following insights:

i) We prove that in the NDS regime, the loss of routing jobs on arrival as opposed to delaying until an idle server becomes available (called the central queue model) is bounded by 14%. Our experiments show for as few as 4 server the worst case loss is indeed between 12-16%. Therefore NDS faithfully models the behavior of finite systems despite being an asymptotic regime.

ii) We prove that an alternative to SQ that has been proposed in the literature, called Idle-Queue-First (IQF), incurs a loss of 100% compared to central queue dispatcher. However a slight modification to IQF, which we call Idle-One-First (I1F) reaps all the benefits of SQ.

iii) In ongoing work we show that the mean completion time under SQ load balancing with Processor Sharing servers is indeed insensitive under NDS, as observed in our experiments, while workload based load balancing is not. This result clinches the deal for NDS regime, as NDS is able to separate SQ from workload based load balancing while neither of the two asymptotic regimes proposed in the literature does.

This research also fits nicely with the lens of distribution oblivious control – our results suggest that SQ (which is oblivious to the size of the jobs as well as to the distribution of the job sizes) only loses approximately 14% with respect to the optimal size-aware load balancing policy!

IV. Future Directions

Applications to Large Scale Systems

My research so far has focused on deriving mathematical insights into the structure and robustness of control policies for stochastic service systems. I increasingly find a disconnect between the research and the needs of practitioners driven by two gaps:

1. *The models studied are extreme:* While the models used by the stochastic processes community are extreme in that they impose strong statistical regularity assumptions, the models used in theoretical computer science also tend to be extreme because they do not impose any assumptions. Recently many models within this spectrum have emerged – random order, or semi-random models, online algorithms with predictions, robust optimization. When considering strategic agents, a lot of the research assumes perfectly rational agents and that the system designed knows the

distribution of utilities for the agents. Here again there are emerging trends in studying prior-free and simple mechanisms, and quantifying the efficiency loss. Going forward my goal is to continue working towards a practically useful theory of online algorithms and stochastic control. Towards that goal, I designed a PhD course titled *Online Optimization and Decision Making under Uncertainty* to expose our graduate students to the diverse models used to study these problems.

2. *Lack of computational tools:* Faced with the problem of controlling large scale systems, practitioners invariably resort to heuristics, or deterministic optimization problems based on deterministic forecasts of the future. A second goal of my research will be to develop efficient computational tools for high-dimensional MDPs which utilize queueing and control theory, algorithms for learning under non-stationary data, simulation models, approximate dynamic programming and representation learning. A prototypical example of such high-dimensional MDPs is spatial inventory management – such as managing inventory among a network of warehouses and retailers, inventory of trailers in a freight network, if cars/bikes/scooters in ride/bike/scooter-sharing. Here the geographic proximity of nodes creates natural spillover and correlation among demand at the nodes, as well as tradeoff between serving from the closest location which is low on inventory, versus a farther location with ample inventory.

Interface of OR/AI with Public Policy and Public Health

A second agenda I would like to pursue in the near future is to explore the intersection of public policy and operations. A project that is close to my heart is to use tools from operations research and AI to improve access to mental health care and lower the barriers for their access. While in our field we devote a lot of attention to resources such as online ads, airplane seats, medical equipment, we do not devote enough attention to human capital that is wasted via poor access to education as well mental healthcare. [2] Close to our own territory, there have been reports of increasing mental health problems among college students, [3] and among PhD students. [4] On policy front, the U.S. Preventive Services Task Force recently published their recommendation for mental health screenings for children. [5] There have also been interesting developments in using AI tools such as chatbots [6] and games [7] as potential gateway treatments. There is much more that needs to be done, e.g., better and more objective diagnosis than a simple question from a health care provider asking if you have been depressed lately; an evidence-based approach to treatment which includes community support, cognitive behavioral therapy, medication, AI tools; more resources devoted to training mental health professionals; AI nudges delivered to improve adherence to treatment.

[2] Insel, Thomas R. Healing : Our Path from Mental Illness to Mental Health. New York : Penguin Press, 2022.

[3] https://www.apmreports.org/episode/2021/08/19/under-pressure-the-college-mental-health-crisis

[4] Satinsky, E.N., Kimura, T., Kiang, M.V. et al. Systematic review and meta-analysis of depression, anxiety, and suicidal ideation among Ph.D. students. Sci Rep 11, 14370 (2021)

[5] https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/screening-anxiety-children-adolescents

[6] https://www.verywellhealth.com/using-artificial-intelligence-for-mental-health-4144239

[7] https://www.prnewswire.com/news-releases/deepwell-dtx-unveils-winners-for-inaugural-mental-health-game-jam-301578977.html

PUBLICATIONS

JOURNAL

[1] V. Gupta. Greedy Algorithm for Multiway Matching with Bounded Regret. *Operations Research*. Forthcoming, 2022.

[2] Y. Luo, V. Gupta, and M. Kolar. Dynamic Regret Minimization for Control of Non-stationary Linear Dynamical Systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*. 6(1), 2022.

[3] V. Gupta, and J. Zhang. Diffusion Approximation and Optimal Control for State-dependent Limited Processor Sharing Queues. *Stochastic Systems*. 12(2), 2022.

[4] P. Afeche, R. Caldentey and V. Gupta. On the Optimal Design of Bipartite Matching Queueing Systems. *Operations Research*. 70(1), 2022.

[5] V. Gupta and A. Radovanovic. Interior-point Based Online Stochastic Bin Packing. *Operations Research*. 68(5), 2020.

[6] M. Yu, V. Gupta and M. Kolar. Estimation of a Low-rank Topic-Based Model for Information Cascades. *Journal of Machine Learning Research*. 21(71), 2020.

[7] V. Gupta, B. Moseley, M. Uetz and Q. Xie. Greed Works - Online Algorithms For Unrelated Machine Stochastic Scheduling. *Mathematics of OR*. 45(2), 2020.

[8] V. Gupta and N. Walton. Load Balancing in the Non-Degenerate Slowdown Regime. *Operations Research*. 67(1), 2019.

[9] L. Yang, L. Debo, and V. Gupta. Search among Queues under Quality Differentiation. *Management Science*. 65(8), 2019.

[10] L. Yang, L. Debo, and V. Gupta. Trading Time in a Congested Environment. *Management Science*. 63(7), 2016.

[11] A. Busic, V. Gupta, and J. Mairesse. Stability of the Bipartite Matching Model. *Advances in Applied Probability*. 45(2), 2013.

[12] V. Gupta and T. Osogami. On Markov-Krein characterization of the mean sojourn time in $M/G/K$ and other queueing systems. *Queueing Systems*, 68(3-4):339–352, 2011.

[13] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. A. Kozuch. Optimality analysis of energy-performance trade-off for server farm management. *Performance Evaluation*, 67(11):1155–1171, 2010.

[14] V. Gupta, M. Burroughs, and M. Harchol-Balter. Analysis of scheduling policies under correlated job sizes. *Performance Evaluation*, 67(11):996–1013, 2010.

[15] V. Gupta, J. Dai, M. Harchol-Balter, and B. Zwart. On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Systems*, 64(1):5–48, 2010.

[16] M. Vojnovic, V. Gupta, T. Karagiannis, and C. Gkantsidis. Sampling Strategies for Epidemic-style Information Dissemination. *IEEE Transactions on Networking.* 18(4), 2010.

[17] V. Gupta, M. H. Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007.

CONFERENCE & WORKSHOP

[18] V. Gupta, R. Krishnaswamy, S. Sandeep, and J. Sundaresan. Look Before, Before You Leap: Online Vector Load Balancing with Few Reassignments. *Innovations in Theoretical Computer Science (ITCS)*, 2023.

[19] Z. Gao, J. Birge, and V. Gupta. Approximation Schemes for Multiperiod Binary Knapsack Problems. *International Computer Science Symposium in Russia,* 2021.

[20] V. Gupta, R. Krishnaswamy and S. Sandeep. Permutation Strikes Back: The Power of Recourse in Online Metric Matching. *APPROX/RANDOM* 2020.

[21] C. Zhang, V. Gupta, and A. Chien. Information Models: Creating and Preserving Value in Volatile Cloud Resources. *International Conference on Cloud Engineering (IC2E),* 2019.

[22] M. Yu, V. Gupta, and M. Kolar. Learning Influence-Receptivity Network Structure with Guarantees. *AISTATS*, 2019.

[23] M. Yu, V. Gupta, and M. Kolar. An Influence-Receptivity Model for Topic Based Information Cascades. *ICDM* 2017.

[24] M. Yu, V. Gupta and M. Kolar. Statistical Inference for Pairwise Graphical Models Using Score Matching. *NeurIPS* 2016.

[25] S. Borst, V. Gupta, and A. Walid. Distributed Caching Algorithms for Content Distribution Networks. *IEEE INFOCOM.* 2010.

[26] H. Amur, J. Cipar, V. Gupta, M. Kozuch, G. Ganger, and K. Schwan. Robust and Flexible Power-proportional Storage. *Symposium on Cloud Computing (SOCC).* 2010.

[27] V. Gupta and M. Harchol-Balter. Self-Adaptive Admission Control Policies for Resource-Sharing Systems. *ACM SIGMETRICS/Performance.* 2009.

[28] V. Gupta and P. Harrison. Fluid Level in a Reservoir with an On-Off Source. *ACM SIGMETRICS Performance Evaluation Review.* 36(2), 2007.

[29] V. Gupta Finding the Optimal Quantum Size: Sensitivity Analysis of the $M/G/1$ Round-Robin Queue. *ACM SIGMETRICS Performance Evaluation Review.* 36(2), 2007.

[30] V. Gupta, M. Harchol-Balter, A. Scheller-Wolf, and U. Yechiali. Fundamental Characteristics of Queues with Fluctuating Load. *ACM SIGMETRICS/Performance*. 2006.

WORKING PAPERS

[31] R. Caldentey, V. Gupta and L. Hillas. Designing Service Menus for Bipartite Queueing Systems with Strategic Users. *Operations Research*. (Under revision)

[32] T. Akturk, O. Candogan and V. Gupta. Managing Resources for Shared Micromobility: Approximate Optimality in Large-Scale Systems. Under submission

[33] V. Gupta. Reneging and Balking in Resource Sharing Systems. Available at SSRN: `https://ssrn.com/abstract=4158671`, 2022.

[34] L. Debo, P. Enders, A. Gandhi, V. Gupta, M. Harchol-Balter, and A. Scheller-Wolf. Inducing Optimal Scheduling with Selfish Users.

WORK IN PROGRESS

[35] P. Afeche, V. Gupta and A. Mani. Service differentiation in Spatial Queueing Systems.

[36] V. Gupta. Dynamic Auctions for Truckload Marketplaces.

[37] N. Thakurele and V. Gupta. Trade-offs in Preferential Access Auctions.